Measuring Retrieval Effectiveness: A New Proposal and a First Experimental Validation

Vincenzo Della Mea and Stefano Mizzaro
Department of Mathematics and Computer Science, University of Udine, Via delle Scienze, 206, I 33100 Udine, Italy. E-mail: {dellamea, mizzaro}@dimi.uniud.it

Most common effectiveness measures for information retrieval systems are based on the assumptions of binary relevance (either a document is relevant to a given query or it is not) and binary retrieval (either a document is retrieved or it is not). In this article, these assumptions are questioned, and a new measure named ADM (average distance measure) is proposed, discussed from a conceptual point of view, and experimentally validated on Text Retrieval Conference (TREC) data. Both conceptual analysis and experimental evidence demonstrate ADM's adequacy in measuring the effectiveness of information retrieval systems. Some potential problems about precision and recall are also highlighted and discussed.

Introduction

In the information retrieval (IR) field, most common measures of the effectiveness of an information retrieval system (IRS) are based on binary relevance (either a document is relevant to a given query or it is not) and binary retrieval (either a document is retrieved or it is not). These assumptions can, and need to, be questioned; relevance might be not binary, and IRSs usually rank the retrieved documents and, sometimes, show their weights (e.g., all the Web search engines, let alone the vector space–based IR system existing since the 1970s).

In this article, which revises from a conceptual point of view and extends with some experimental data some previous work (Mizzaro, 2001), we define the average distance measure (ADM), a new measure of retrieval effectiveness, and discuss its adequacy both by means of a conceptual analysis and by presenting some experimental data.

The article is structured as follows. In the next section, we briefly survey the issue of retrieval evaluation, emphasizing the underlying assumptions of dichotomous conception of both relevance and retrieval. In the following section, we define ADM, a new measure of retrieval effectiveness based on a continuous view of relevance and retrieval. In the "Adequacy of ADM" section, we show how ADM is adequate for measuring the effectiveness of IRSs, and how it leads us to both highlight and overcome some problems inherent in the effectiveness measures usually adopted in retrieval evaluation, namely, precision and recall. In the "Experimental results" section, we present some experimental evidence supporting the adequacy of ADM. The last section concludes the article and sketches some future developments.

Measuring Retrieval Effectiveness

Some Problems in Measuring IR Effectiveness

Traditionally, given an information need and the corresponding query, the database of documents is partitioned in two ways, as it is graphically represented in Fig. 1(a), adapted from Salton & McGill (1984): (i) between relevant and not relevant items, and (ii) between retrieved and not retrieved items. A reason for this approach is historical: The first IRSs were boolean, and they indeed provided a clear-cut distinction between retrieved and nonretrieved documents. From that, it is (and probably has been) a small step to assume that relevance is binary as well, and, given the binary conceptions of relevant and retrieved documents, the definition of precision (i.e., the proportion of retrieved documents that are relevant) and recall (i.e., the proportion of relevant documents that are retrieved) is (has been) a logical consequence.

Actually, the two underlying assumptions (binary relevance and binary retrieval) have been questioned for a long time. On the one side, after the first IRSs based on the vector space and probabilistic models, it has been clear that an IRS does not "either retrieve or not retrieve a document," but rather ranks all the documents in the database on the basis of some system-assigned weight. This is widely known today, since everybody has experienced some search engine. On the other side, the long record of research on relevance (Mizzaro, 1997) indicates

Received January 6, 2003; revised September 23, 2003; accepted October 27, 2003

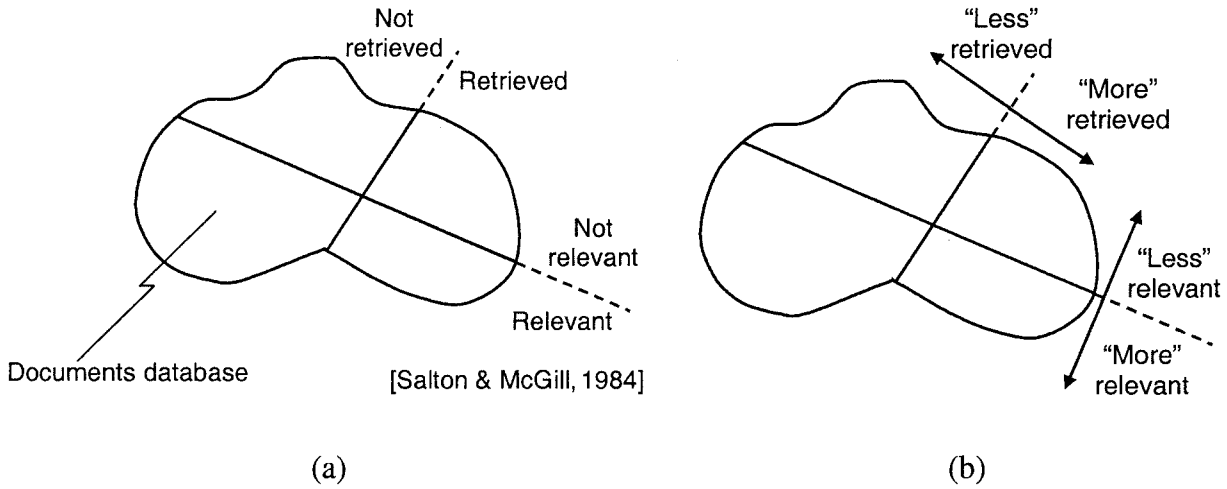

FIG. 1. From binary relevance and retrieval to continuous relevance and retrieval.

that relevance is not binary and binary judgments do not seem the most adequate method of expression (Bruce, 1994; Eisenberg, 1986; Eisenberg, 1988; Janes, 1991b; Janes, 1994; Janes & McKinney, 1992).

Indeed, some measures that go beyond the binary relevance-binary retrieval view have been proposed; most of them are well known (and described in standard IR textbooks, see, e.g., Korfhage, 1997, Ch. 8, Salton & McGill, 1984, Ch. 5; van Rijsbergen, 1979, Ch. 7;), and are sometimes used. Ignoring the other measures based on the same assumptions (i.e., fallout, generality factor, E-measure, mean average precision, and so on), by discarding the binary retrieval assumption we obtain measures based on the ranking induced by the IRS (i.e., normalized precision and recall, expected search length) or even on a continuous measure provided by the IRS (e.g., Swets's E-measure). If we also discard the binary relevance assumption, we obtain measures that can be classified in three groups:

- Measures based on categories of relevance and the rank produced by the IRS, e.g., Ranked Half Life (Borlund & Ingwersen, 1998) or Discounted Cumulative Gain (Järvelin & Kekäläinen, 2000).
- Measures that compare the ranking induced by the IRS with the ideal one, e.g., ndpm (Yao, 1995) or usefulness measure (Frei & Schauble, 1991).
- Measures that evaluate the IR effectiveness using continuous values of relevance and retrieval, like the sliding ratio.

However, precision and recall have survived all these discussions, and are still widely used as the standard measures of IR evaluation. The standard practice today is still to evaluate IRSs by precision and recall, and, therefore, on the basis of the binary relevance and retrieval assumptions: in IR evaluation, often (if not usually) IRSs are meant to either retrieve or not retrieve a document, and human relevance judgments are dichotomous ones. The well-known Text Retrieval Conference (TREC) experiment series is an example of this approach, even if in TREC the binary retrieval view is in some way smoothed by the procedure requiring 1,000 ranked documents being returned by each system, and the adopted effectiveness measures are derivations of precision and recall.

The standard practice is so deeply rooted that, even when human relevance judgments are not dichotomous (i.e., they are expressed either by means of a scale of categories or on a continuum), precision and recall often cause a "binarization" of the judgments. For example, it is often assumed that, on a three-level scale (i.e., nonrelevant, partially relevant, and relevant), the partially relevant items collapse into relevant ones (Schamber, 1994; Spink, Greisdorf, & Bateman, 1998) and/or (less frequently) into nonrelevant ones (Voorhees, 2001); also continuous judgments are binarized (Eisenberg, 1988; Eisenberg & Hu, 1987; Rorvig, 1988; Schamber, 1994). Even if there is some experimental motivation for preferring "relevant" to "nonrelevant" when collapsing "partially relevant" (Eisenberg, 1986; Eisenberg & Hu, 1987; Janes, 1991a), there is absolutely no reason for binarizing the relevance judgments (apart from being able to compute precision and recall). Similar phenomena do happen on the retrieval side too, where it is common to speak of "the retrieved documents," or of "the first page of documents retrieved by the search engine X." Moreover, the error rates for commonly used measures are far from being negligible, so that, for a reliable IR evaluation experiment, 50 queries are needed, and for having a significant difference between two IRSs, a 10% difference in IR performances is needed (Buckley & Voorhees, 2000; Buckley & Voorhees, 2002; Sparck Jones, 1974). Finally, the measures obtained starting from artificially binarized figures are eventually averaged on many data, thus obtaining the rather peculiar result of continuous values.

Therefore, the IR field is in a curious situation: on the one side, we have a "conceptual" standard, since almost everybody agrees that relevance and retrieval are matter of degree (three or more categories, if not a continuum); on the other side, we have an old habit, the "precision-recall old standard," which relies on the assumption of binary relevance and retrieval. This

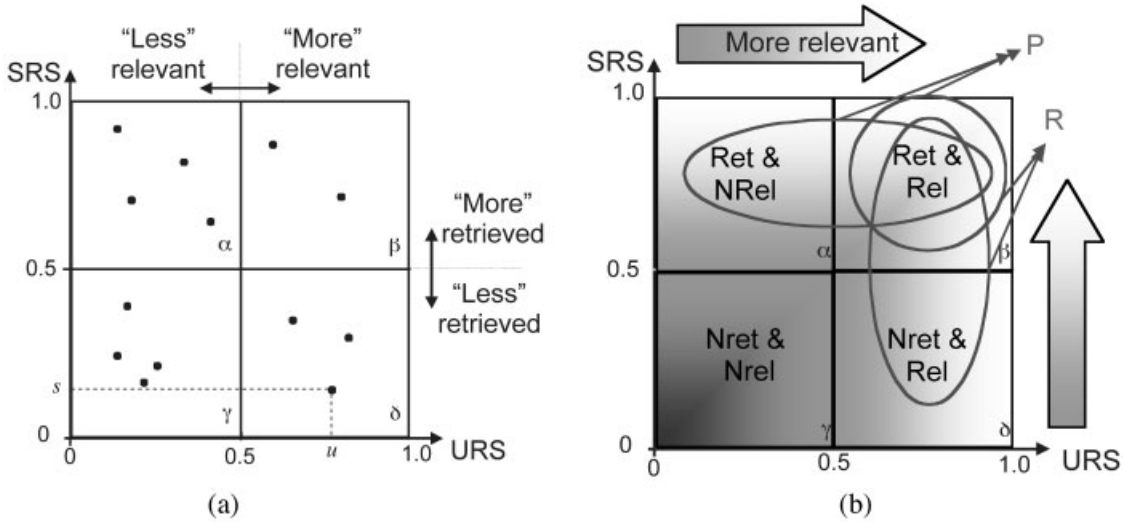

FIG. 2. The URS–SRS plane.

situation has the consequence that most of the evaluation experiments disregard the "conceptual" standard, thus hindering IR development and evaluation.

This impasse is dangerous since researchers risk: (i) evaluating in the wrong way the IRSs they are developing; (ii) developing "wrong" IRSs (i.e., IRSs that are evaluated as effective by the wrong measures but are not so effective); and (iii) exerting more effort than needed for evaluating IR effectiveness.

We propose a novel approach.

From Binary to Continuous Relevance and Retrieval

We generalize Fig. 1(a) as shown in Fig. 1(b): in place of two clear-cut partitions, we have gradients of relevance and retrieval. By going one step further, we make explicit the two figures that measure relevance and retrieval. As far as relevance is concerned, we define the user relevance score (URS) as a value in the [0,1] range that measures the real (i.e., user-determined) relevance of a document with respect to an information need. URS assumes the maximum value (i.e., 1) for "totally relevant" documents, a 0 value for "totally nonrelevant" items, and intermediate values for documents with various degrees of "partial" relevance. Conversely, the retrieval measure is named *system relevance score* (SRS): the score of the relevance of a document to a query given by the IRS. SRS has the same behavior as URS: it is in the [0,1] range, and 1 is its maximum value. Boolean IRSs give either SRS = 0 or SRS = 1.[1]

On the basis of the definitions of URS and SRS, we can slightly change the graphical representation in Fig. 1(b), obtaining Fig. 2(a), which shows a URS-SRS plane in which each document is a point with its own URS and SRS values (in the figure, u and s are these values for one document, represented by the point in the lower right corner).

This representation emphasizes how the dichotomies relevant-nonrelevant and retrieved-nonretrieved correspond to the (somewhat artificial and hardly justifiable) choice of two thresholds on the SRS and URS values. Fig. 2(b) is yet another representation of the same scenario, with the color shading representing the two gradients. In this figure, the ellipses show which documents concur to determining precision (P) and recall (R). Indeed, on the basis of Fig. 2, one might define precision, recall, fallout, and generality factor in the following way:

$$P = \frac{|\alpha|}{|\alpha| + |\beta|}, \quad R = \frac{|\beta|}{|\beta| + |\delta|},$$

$$F = \frac{|\alpha|}{|\alpha| + |\gamma|}, \quad G = \frac{|\beta| + |\delta|}{|\alpha| + |\beta| + |\gamma| + |\delta|},$$

where $|\alpha|$, $|\beta|$, $|\gamma|$, and $|\delta|$ are the numbers of points, i.e., documents, in the α, β, γ, and δ sectors, respectively. Of course, one might choose two thresholds on each axis and single out, in this way, nine regions or, in general, n thresholds and $(n+1)^2$ regions. However, an even more general case is the continuous one that we exploit to define a new measure of retrieval effectiveness, as shown in the next section.

Of course, there is the problem of collecting URS and SRS values. On the one hand, obtaining URSs seems feasible in various ways. One could simply use standard dichotomous—or category—relevance judgments. By averaging several such judgments by different judges on the same document-

[1]SRS is similar to retrieval status value (RSV) (Bookstein, 1979), but there is a difference: RSVs are used only to rank the documents and, therefore, any transformation of a RSV distribution that preserves the ranking is another equivalent RSV distribution. This is not the case for SRS, as we will discuss later in this article. The difference stems from the underlying notion of relevance: RSV is based on binary relevance, SRS on continuous.

query pair, a continuous value is obtained. Or, one could use magnitude-estimation techniques: line length and force hand grip have been used in the past rather effectively (Bruce, 1994; Janes, 1991b; Janes, 1994; Rorvig, 1988).

On the other hand, to have IRSs computing true SRSs requires new IR models and a new approach to IRS implementation. At a first stage, one might think of using probabilistic and vector space IRSs, but it is important to note that both the estimation of the probability of relevance given by a probabilistic IRS and the distance between the query and document vectors given by a vector space IRS are not estimations of the amount of relevance of a document to a query. To obtain such an estimation, new IRSs based on new IR models are needed.

A last important issue that we mention is the apparent arbitrary nature of URSs and SRSs. Even if URSs might seem arbitrary at first, they turn out to be not arbitrary at all if they can be elicited reliably and consistently from human relevance assessors. And the above cited studies on magnitude-estimation techniques (Bruce, 1994; Janes, 1991b; Janes, 1994; Rorvig, 1988) are some first positive results in this direction. Now, if URSs are not arbitrary, SRSs turn out not to be arbitrary too: the correct SRS for a document with respect to a query is the URS of that document for that query. This natural observation leads to the evaluation measure proposed in the next section.

The Average Distance Measure

We propose a new retrieval effectiveness measure, named *average distance measure* (ADM), based on the average distance, or difference, between URSs (the actual relevance of documents) and SRSs (their estimates by the IRS). To have 0 as the minimum value, and 1 as the maximum value (as is common in IR evaluation), we subtract the average distance from 1. In a more formal way, for a given query q, we define two relevance weights for each document d_i in the database D: the SRS for d_i with respect to q (denoted by $SRS_q(d_i)$), and the URS for d_i with respect to q $URS_q(d_i)$). ADM for the query q is then defined as the average distance between $SRS_q(d_i)$ and $URS_q(d_i)$:

$$ADM_q = 1 - \frac{\sum_{d_i \epsilon D} |SRS_q(d_i) - URS_q(d_i)|}{|D|} \qquad (1)$$

(where the denominator is the number of documents in the database D). ADM_q is in the [0,1] range, with 0 representing the worst performance. By averaging ADM_q on some queries, we obtain ADM, a measure of IR effectiveness.

We can graphically understand ADM in the following way. Let us assign to each document in the database its own SRS and URS values (in the [0,1] range) and plot these values on a standard Cartesian diagram in the $[0,1]^2$ square (see Fig. 3). Each document is therefore a point in the URS-SRS plane; the closer the point to the ideal SRS = URS line (the dotted line in the figure), the better the

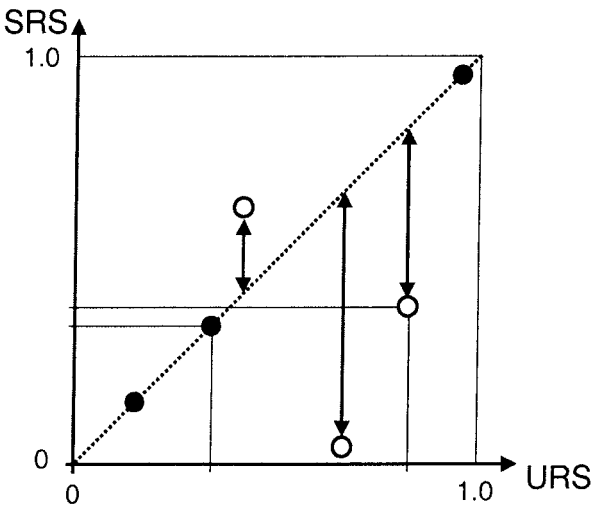

FIG. 3. Graphical representation of ADM.

estimate by the IRS (the points on the line are represented by filled circles in figure). The last thing we need to define is the distance between a point and the ideal line. Since the URS value is predefined and cannot be changed as a result of the retrieval of a document,[2] the distance is not the standard distance between a point and a line (i.e., the length of an orthogonal line from the point to the line), but the distance between the point representing the document and the point on the line with the same abscissa (represented by the arrows in figure). This is the definition used in Eq. (1).

Let's see an example. Table 1 shows three hypothetical documents, with their URSs and the corresponding SRSs for three different IRSs. The last four columns of the table contain the values for precision, recall, E-measure (defined here as the mean between precision and recall), and ADM for the three IRSs, under the assumption that both the thresholds—between relevant and nonrelevant and between retrieved and nonretrieved—are 0.5 (values ≥ 0.5 are bold in the table). See also Fig. 4, where circles are IRS1 points, crosses are IRS2 points, and squares are IRS3 points.

Let us briefly analyze this example (more detailed discussion about ADM follows in the next section). System IRS1 performs constantly better than IRS2 (each circle is closer to the ideal SRS = URS line than the corresponding cross); this is not reflected in the values of the three classical measures, whereas ADM captures the difference in effectiveness. Systems IRS1 (circles) and IRS3 (squares) are more difficult to compare, since IRS3 performs better than IRS1 on all but one of the documents (d_3), but on d_3 the SRS by IRS3 is really wrong. Precision, recall, and E-measure

[2]In this article, we do not take into account the subjective and dynamic nature of relevance (Mizzaro, 1998; Schamber, Eisenberg, & Nilan, 1990), and we assume that the user is able to determine the "real" relevance value. However, our results can be extended in a straightforward way to the more general case of the user view of relevance.

for IRS1 and IRS3 do not differ, whereas there is a difference in the two ADM values.[3]

Also, specialized forms of ADM can be defined. ADM can be specialized into an $ADM^{(2)}_{(2)}$ measure to handle the binary relevance-binary retrieval view. In this case, all the points in the URS-SRS plane turn out to be in either (0,0), (0,1), (1,0), or (1,1) and, therefore, the distances from the ideal line are either 0 or 1. When it is possible to associate a numeric value to ordinal categories, it is also straightforward to define: $ADM^{(M)}_{(N)}$, based on N categories of relevance and M categories of retrieval (i.e., URSs assume one of N values, and SRSs assume one of M values), $ADM^{(M)}$ (with M categories of retrieval and continuous relevance), and $ADM_{(N)}$ (with N categories of relevance and continuous retrieval).[4]

Finally, ADM can be tuned in a very simple way, just by selecting the sample of documents used for its computation. For example, if only the most relevant documents are used, one measures the accuracy of the IRS in estimating the user relevance on the highly relevant documents only, and this seems a very important measure from the user point of view (Järvelin & Kekäläinen, 2000; Voorhees, 2001).

Adequacy of ADM

In this section, we show, from a conceptual point of view, how ADM is adequate for measuring the effectiveness of IRSs, in some respect even more adequate than classical precision and recall.

ADM satisfies the classical four desirable properties proposed by Swets (1967) and reported also in van Rijsbergen (1979, Ch. 7): it measures the effectiveness only, isolating it from efficiency and cost; it expresses the discrimination power of IRSs, independently of any acceptance criterion employed; it is a single number; and it allows complete ordering of different performances. Of course, ADM is not the only IR effectiveness measure that satisfies these properties (e.g., the E-measure does), nor do these four properties guarantee that ADM is a good measure, since they are necessary and not sufficient conditions.

ADM adequacy is clearly shown when we compare it with other IR effectiveness measures usually adopted in retrieval evaluation. What follows concerns mainly precision and recall, but it can be generalized to other measures as well. This comparison, besides being useful for discussing ADM adequacy, will also lead us to reconsider the classical effectiveness measures by highlighting their intrinsic limitations.

We can compare ADM with precision and recall on the basis of Fig. 2. ADM is, in some sense, more general, since:

- Precision and recall take into account the documents in some of the four sectors only (e.g., precision is based on sectors α and β only). If, in Fig. 2(a), some points were added to the γ sector, either close to the ideal line or far from it, neither precision nor recall would be affected. However, if the points were close to (far from) the ideal SRS = URS line, this would mean that the IRS has correctly (wrongly) estimated the relevance of the corresponding documents, and therefore its effectiveness measure should increase (decrease). This is also a justification for preferring the recall-fallout pair to the recall-precision one: the former covers the whole $[0,1]^2$ sector, while the latter covers just 75% of it (α, β, and δ), and the 75% with less documents in it, since most of them will be in the γ sector (in general, given a query, most of the documents are neither relevant nor retrieved).
- Precision and recall do not use the full-fledged distance from the ideal line used in Eq. 1, since all the documents within each sector (α, β, γ, and δ) are considered as equivalent (the distance used is 0 if the document is in sector β or γ, 1 if the document is in sector α or δ: the same limitation of $ADM^{(2)}_{(2)}$).

This comparison between ADM on the one side and precision and recall on the other shows how rough precision and recall are. The second point above also reveals two further problems. First, precision and recall are highly (too) sensitive to the thresholds chosen and to the documents close to the borders between sectors. Fig. 5(a) shows how three documents might be judged by three hypothetical IRSs (circles represent IRS1, crosses IRS2, and squares IRS3). Clearly, the three systems are extremely similar, or at least evaluate the three documents in very similar ways. However, the values for precision, recall, E-measure (assuming again that the two thresholds—between relevant and nonrelevant and between retrieved and nonretrieved—are 0.5), and ADM (Table 2(a)) show that classical measures are rather different, whereas ADM is more stable.

TABLE 1. An example.

Docs.	d_1	d_2	d_3	P	R	E	ADM
URS	**0.8**	0.4	0.1				
IRS1○	**0.9**	**0.5**	0.2	0.5	1	0.75	0.9
IRS2×	**1.0**	**0.6**	0.3	0.5	1	0.75	0.8
IRS3□	**0.8**	0.4	**1.0**	0.5	1	0.75	0.7

[3]The SRSs given by IRS1 and IRS2 lead to the same ranking of the three documents. Therefore, they are equivalent if interpreted as RSV (see Footnote 1). However, if the IRSs have the aim of finding the best approximation of URSs, IRS1 is more effective than IRS2.

[4]The assignment of numerical value to ordinal categories can present subtle problems. As a matter of fact, the "linear scale assumption," (i.e., the naïve assumption that the categories correspond to equally distant URS—or SRS—values) can be easily questioned. This can be seen by means of a simple example. If we have three categories labeled "relevant," "partially relevant," and "not relevant," it seems rather natural to give them 1, 0.5, and 0 values. But why should this assignment be preferred to, say, the 1, 0.6, 0 choice? Moreover, the symmetry considerations that might help in this case do not hold if the labels of the three categories are "highly relevant," "relevant," and "not relevant," for which the values are even more arbitrary. Anyway, any solution seems better than collapsing the intermediate relevance categories into "relevant" or "not relevant." this latter choice is the one with the highest error rate. We will briefly come back to this issue (which has been brought to our attention by Steve Robertson) in the last section of this article.

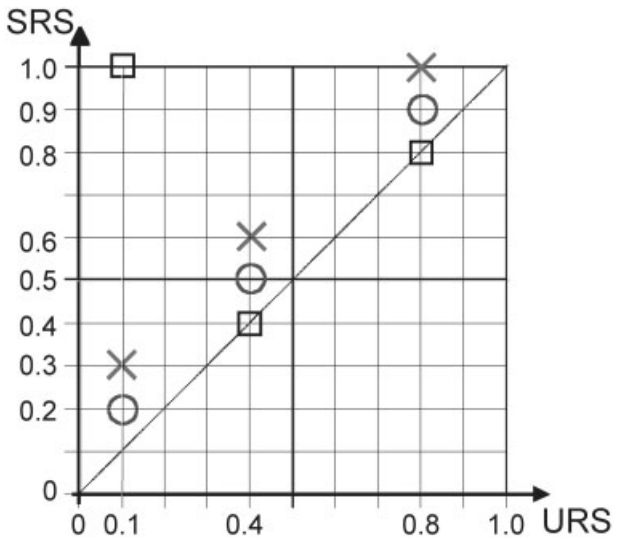

FIG. 4. Graphical representation of the example in Table 1.

The second problem is that precision and recall are not sensitive enough to important differences between systems. Fig. 5(b) shows how two documents might be judged by two hypothetical systems (circles stand for IRS1, crosses for IRS2). Clearly, the two systems evaluate the two documents in rather different ways. The values for precision, recall, E-measure, and ADM (Table 2(b)) show that classical measures are completely unable to grasp the difference, whereas ADM clearly differentiates the effectiveness of the two systems.

Therefore, the two problems about precision and recall are: (1) small differences in the SRS can lead to very different precision, recall, and E-measure figures, whereas small differences do not affect ADM and (2) big differences in SRS can lead to very similar (even identical) precision, recall, and E-measure figures, whereas big differences do affect ADM.

TABLE 2. Effectiveness measures for Figs. 5(a) and 5(b).

	P	R	E	ADM
IRS1○	0.67	1	0.84	0.83
IRS2×	1	0.5	0.75	0.83
IRS3□	0.5	0.5	0.5	0.826

(a)

	P	R	E	ADM
IRS1○	1	1	1	1
IRS2×	1	1	1	0.5

(b)

Both problems are relieved in real IRS evaluation, since precision and recall figures are obtained by averaging many queries retrieving many documents. However, they might be one reason for the high variation of precision and recall among different queries (often higher than the variation among different IRSs) (Harter, 1996). Moreover, looking at it from a different perspective, by using ADM in place of precision and recall, information-retrieval experiments may be carried out on smaller data sets (less queries), and the effectiveness for queries with very few relevant documents is measured in a more reliable way.

Both problems depend on the thresholds on SRS and URS. The second one, however, has a further component: the equal status given to documents within each sector (α, β, γ, and δ; see Fig. 2(a)) in the calculation of precision and recall. Indeed, it seems unfair to consider all the documents in, say, β sector simply as "retrieved and relevant"; a fairer categorization might be the one shown in Fig. 6(a), where the documents in the brighter area $\alpha 1$ (closer to the ideal

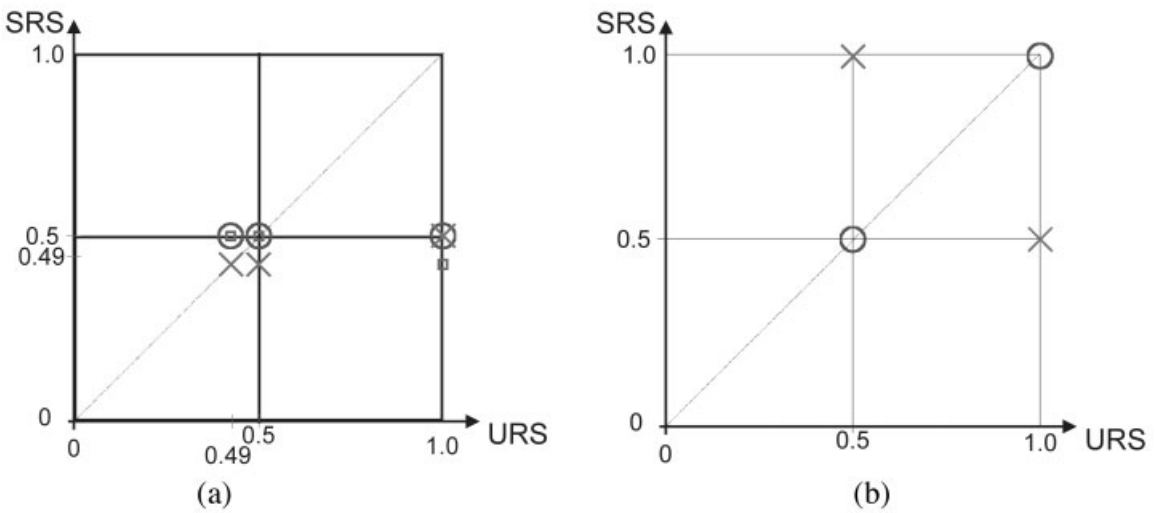

FIG. 5. Small (a) and big (b) differences in SRS values.

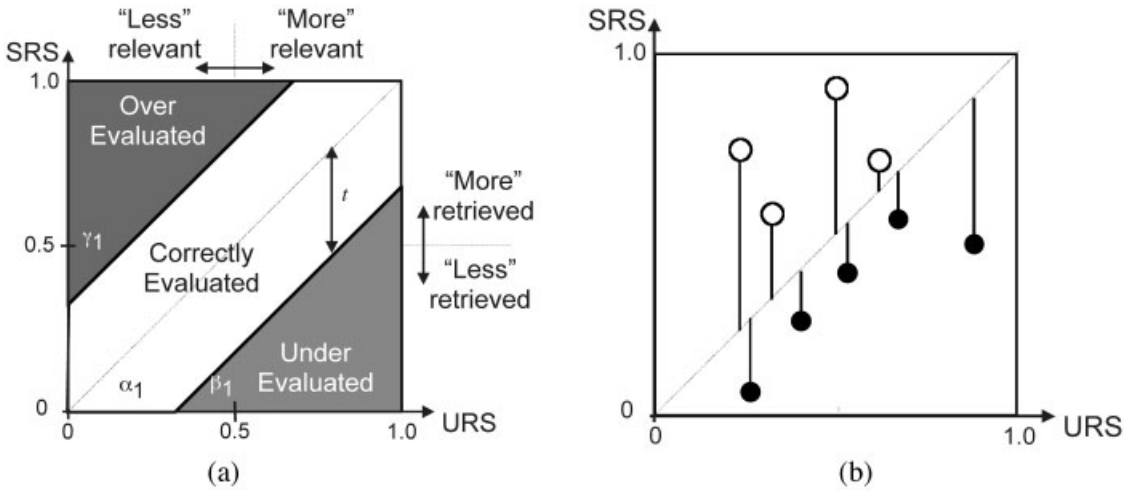

FIG. 6. A better categorization than the classical one in Fig. 2(b).

line) are considered as correctly evaluated (their URS-SRS distance is below a given threshold value t in figure), whereas the document in the darker areas $\beta 1$ and $\gamma 1$ are not correctly evaluated. Correct evaluation, of course, leads to higher IR effectiveness.

On the basis of this categorization, one could define two substitutes for precision and recall. Let us start by noticing that recall can be considered an inverse measure of relevance underevaluation (regarding relevant, i.e., 1, as higher than nonrelevant, i.e., 0), since it depends on the number of relevant documents considered not relevant. In the same way, precision is an inverse measure of relevance overevaluation.

Starting from these properties, we may remark that underevaluation can be expressed also as an IRS assigning SRS values that are lower than the URS values (leading to lower recall values): the points (documents) that should be placed in the upper right corner tend to be moved in the lower right corner of the URS-SRS plane, and not retrieved (think, for example, of a document that is relevant, i.e., URS close to 1, but is underevaluated as not relevant, i.e., SRS close to 0). Conversely, overevaluation can be described as an IRS assigning SRS values higher than the URS values, leading to lower precision values, since more documents (points) are retrieved (moved towards the upper zone of the URS-SRS plane).

On the basis of these remarks, two hypothetical measures replacing precision and recall might be defined as:

$$P^* = \frac{|\alpha 1|}{|\alpha 1| + |\gamma 1|}, \quad R^* = \frac{|\alpha 1|}{|\alpha 1| + |\beta 1|}$$

(where, as usual, $|\alpha 1|$ is the number of documents in the $\alpha 1$ sector, and the same for the other sectors). P^* inversely measures the number of overevaluated documents, in the same way as precision inversely characterizes the number of nonrelevant documents retrieved by the IRS. Similarly, R^* inversely measures the number of underevaluated documents (i.e., relevant documents "forgotten" by the IRS).

However, these two measures are threshold-based and can be the subject of exactly the same critiques above presented about precision and recall (though the thresholds are chosen in a more sensible way). More specifically, which value to choose for the threshold t? The ideal zero is not feasible because most of the points will not lie on the SRS = URS line, and any other value is completely arbitrary. To avoid these critiques, and by exploiting the full potential of the continuous ADM measure, we can instead define the following two ADM measures, reflecting the original precision-recall pair: (i) average distance precision (ADP), that is, ADM computed on the overevaluated documents only (i.e., those above the ideal SRS = URS line), and (ii) average distance recall (ADR), that is, ADM computed on the underevaluated documents only. In formulae, for each query q:

$$ADP_q = 1 - \frac{\sum_{d_i \epsilon DO} |SRS_q(d_i) - URS_q(d_i)|}{|D|},$$

$$ADR_q = 1 - \frac{\sum_{d_i \epsilon DU} |SRS_q(d_i) - URS_q(d_i)|}{|D|}$$

where DO and DU are the sets of all the over- and underevaluated documents, respectively, and the average distance is subtracted from 1 to have 1 as the value of higher effectiveness). Their graphical representation is shown in Fig. 6(b): ADP_q is the ADM computed only on the white points above the ideal SRS = URS line, whereas ADR_q is ADM computed only on the black points below the SRS = URS line.

ADP and ADR are continuous versions of precision and recall, respectively. Moreover, with these definitions, we

have the nice property that, for each query q, $ADM_q = ADP_q + ADR_q - 1$.

Experimental Results

In the previous section we have shown some advantages that ADM exhibits, from a conceptual point of view, over the standard IR evaluation measures. In this section, we present some experimental data that support ADM adequacy. We tested two main hypotheses:

- ADM is able to measure the IRS effectiveness as usual effectiveness measures do.
- ADM sensitiveness and stability allow to reliably measure IR effectiveness using less data (less topics and less documents) than those usually needed.

To test these hypotheses, we would need a collection of URS and SRS values on some topics and for some IRSs. Unfortunately, no collection of this kind is available, since the publicly available IR test collections are based on a binary view of relevance: URSs are either 0 or 1, and SRSs are computed by the IRSs on the basis of a binary relevance model. To build a continuous relevance test collection from scratch is out of the question, since it would require a twofold, and huge, effort: to have human assessors judging on a continuous scale the relevance of the documents to the queries, and to have newly built, or at least adapted, IRSs that generate SRS values. To be able to do this would also require a lot of further foundational research to understand both which method(s) to adopt to obtain continuous relevance judgments, i.e., URS (e.g., line length magnitude estimation, hand-grip force, averaging of binary or discrete judgments, or something else) and which IR model(s) should be, on the basis of IRSs, capable of generating continuous relevance scores (SRSs). These are open research questions that we indeed plan to address in the future, but are out of the scope of this paper.

On the basis of these considerations, we resorted to using TREC collection to perform our experiments. TREC (http://trec.nist.gov/) is the reference conference series in IR evaluation, and is organized by the National Institute of Standards and Technology (NIST). Its institutional purpose is to support research within the IR community by providing the infrastructure necessary for large-scale evaluation of text-retrieval methodologies. For each TREC conference, NIST provides a (very large) test set of documents, questions, and (binary) relevance assessments to be used by participants to evaluate their own systems.

Let us also remark that using TREC data has positive aspects too. It is a common and well-established way of testing IR systems on somewhat standard data. Since we exploit a commonly recognized test-bed and we start from a somewhat independent source of data, the reliability of our results is somewhat certified (and this would not be true if we used an in-house developed test-bed). Also, exploiting TREC collection allows an experimental comparison between ADM and some standard IR effectiveness measures (Rel-Ret, AvgPrec, R-Prec). This comparison would require more effort with an ad hoc generated test-bed, and this comparison nicely complements the conceptual comparison with P, R, and E-measure in the previous section. Finally, as explained below, we used all TREC queries, whereas some previous studies (e.g., Järvelin & Kekäläinen, 2002; Sormunen, 2002) are based on a somewhat arbitrary selection of TREC queries.

We used data from the ad hoc track (both manual and automatic runs) of TREC-8 (Voorhees & Harman, 2000). These TREC-8 data include the retrieval results of 129 IRSs, each having retrieved 1,000 documents for each of 50 different topics. The first 100 documents of every topic, for every system, are evaluated in order to establish their binary relevance (i.e., relevant/not relevant) by human assessors. Among the evaluations made at TREC-8, we focussed on three effectiveness measures, referred to as *reference measures* in the following: number of relevant documents among the retrieved ones (Rel-Ret), Average Precision (AvgPrec), and R-Precision (R-Prec) (Voorhees & Harman, 2000).

Since TREC-8 data do not contain reliable continuous SRS and URS values, we had to introduce the following simplifications:

- Since weights reported by the systems do not appear to be reliably linked to the effective relevance of the document (they are sometimes inconsistent with the document rankings), we decided to use, as SRS, a normalized measure of the position. Seeing as 1,000 documents were retrieved by each system for each query, the first ranked documents were assigned SRS = 1, the second ranked ones SRS = 0.999, until position 1,000, with value 0.001; zero for all other documents.
- Since the available URS data (TREC's qrels, i.e., human relevance judgments for each query) were binary, we decided for a twofold approach. On the one hand, we used directly the qrels, obtaining URS values hereafter referred to as URS'. On the other hand, to better exploit the potential given by continuous values, we also experimented with another URS, named URS″. URS″ is the weighted average of the qrels and of the average SRSs of the ten best IRSs, according to the formula

$$URS'' = \frac{3 \cdot qrel + 1 \cdot avg(SRS(d_i))}{4}.$$

To select the best systems, we started from those identified in Voorhees & Harman (2000, Figure 7), namely the 5 IRSs in the manual runs with highest mean average precision. To have a higher number of "good" systems, we added to this set those systems, from the manual runs, that exhibit analogous performances (IRSs in the manual runs have better performances than those in the automatic runs).

It is important to understand the positive and negative features of URS′ and URS″. URS″ has the negative property of being based on binary relevance, and therefore of not

exploiting full ADM potential. URS″ is not based on binary relevance values, but it is obtained using the SRSs (of the best systems), therefore introducing a sort of circularity.

We excluded from the experiment all the systems not conforming to the TREC-8 rules on the number of retrieved documents (i.e., all systems retrieving less than 1,000 documents per query), obtaining 109 IRSs (though some of the excluded systems were used to compute URS″). Systems were grouped into three effectiveness categories: *Best* (the group used for calculating URS″, for a total of 7 IRSs after exclusions), *Worst* (all IRSs systematically performing worse than median precision on every topic, for a total of 19), and *Normal* (the 83 remaining systems).

Before presenting the obtained results, let us briefly discuss the general issue of experimental comparison of IR effectiveness measures. To understand which measure is more reliable and sensitive is not an easy task. To compare two (or more) effectiveness measures, statistical correlation seems one of the most frequent choices. For example, Voorhees (2001) used Kendall correlation in the evaluation of IRS (as we do, see below, although in a different setup). Another approach is followed by Buckley and Voorhees (2000), who compared the retrieval performance using the error rate and concluded that a reasonable, though heuristic, notion of difference should be used for comparisons. Järvelin and Kekäläinen (2002) relied on the Friedman test. In order to obtain some more formally founded results, Burgin (1999) proposed the use of the Monte Carlo method as a way for evaluating IRS performance, citing, however, the difficulty in the statistical evaluation of IRS as recognized by van Rijsbergen (1979): "there are no known statistical tests applicable to IR." Even Losee (2000), while aiming at a formal framework for comparing retrieval measures (based on equivalence, equivalent ordering, measure difference function), proposes just an intuitive and graphical evaluation of differences.

First Hypothesis

The average distance measure was computed on each system and topic, for both URS′ and URS″, on all retrieved and relevant documents, and then averaged on all the queries in order to obtain two ADM values for each system (hereafter named ADM′ and ADM″, respectively). Such values were then compared to the reference measures and to the effectiveness categories, to answer the first hypothesis. We did not compute ADM on the whole database since, given the way we assign SRS and URS values, we would obtain that, for a large majority of documents, the distance is zero, because both SRS and URS are zero.

As a first step, we evaluated the nonparametric Kendall's correlation between the rankings induced by the two ADMs and by the three reference measures. We adopted such a correlation measure due to the non-Gaussian distribution of URS and SRS. Table 3 shows the correlation values.

Since correlations between ADMs and reference measures are very high, systems with higher values of ADM (calculated in both ways) are also systems with high values for the three reference measures. As can be seen, correlations between any ADM and any reference measure are of the same order of magnitude as the correlation between reference measures. Furthermore, it should also be noted that correlations between Rel-Ret and both ADMs are higher than correlations between Rel-Ret and both AvgPrec and R-Prec, thus confirming that ADM captures the information related to recall (also expressed by Rel-Ret) better than AvgPrec and R-Prec, more dependent on precision features. Finally, let us also note that ADM′ has a slightly higher correlation than ADM″; therefore, the circularity is not a problem in this case.

TABLE 3. Kendall's correlation between ADM′, ADM″, and reference measures.

	ADM′	ADM″	Rel-Ret	AvgPrec	R-Prec
ADM′					
ADM″	0.874				
Rel-Ret	0.891	0.891			
AvgPrec	0.876	0.857	0.824		
R-Prec	0.844	0.814	0.807	0.902	

After that preliminary evaluation, to better understand the relationship between ADM and reference measures, we graphed the ADMs against Rel-Ret, AvgPrec, and R-Prec, and grouped the systems per effectiveness category. Fig. 7 shows the results:

- A clear association between each of the reference measures with both ADMs is present;
- Best, normal, and worst systems (grouped by means of 95% density ellipses, i.e., the smallest ellipses containing at least 95% points for each category) are identified by the two ADMs in a way similar to that of the three reference measures.

The association between effectiveness category and ADM has been tested using the Kruskal-Wallis test, which identified significant differences among the three categories ($P < 0.0001$), further confirmed on all category pairs by the Mann-Whitney test ($P < 0.0001$ on all pairs).

Figure 8 shows the relationship between the three effectiveness categories and ADM′, ADM″ (boxes represent 90%, 75%, 50%, 25%, and 10% percentiles).

To summarize, the above mentioned results allow us to state that ADM evaluates IRS effectiveness in a way similar to that given by Rel-Ret, AvgPrec, and R-Prec.

Let us emphasize again that the circularity is not a problem: ADM″ relations to the reference measures and to the categories of systems might depend on the way we defined URE″, but ADM″ has no intrinsic circularity, and the results are similar in the two cases. An aspect that needs further investigation is apparent from the last graph of Fig.7. In fact, the rightmost point corresponds to a system that according to the three reference measures is the best one, whereas, according to ADM, it is not so good. This could be

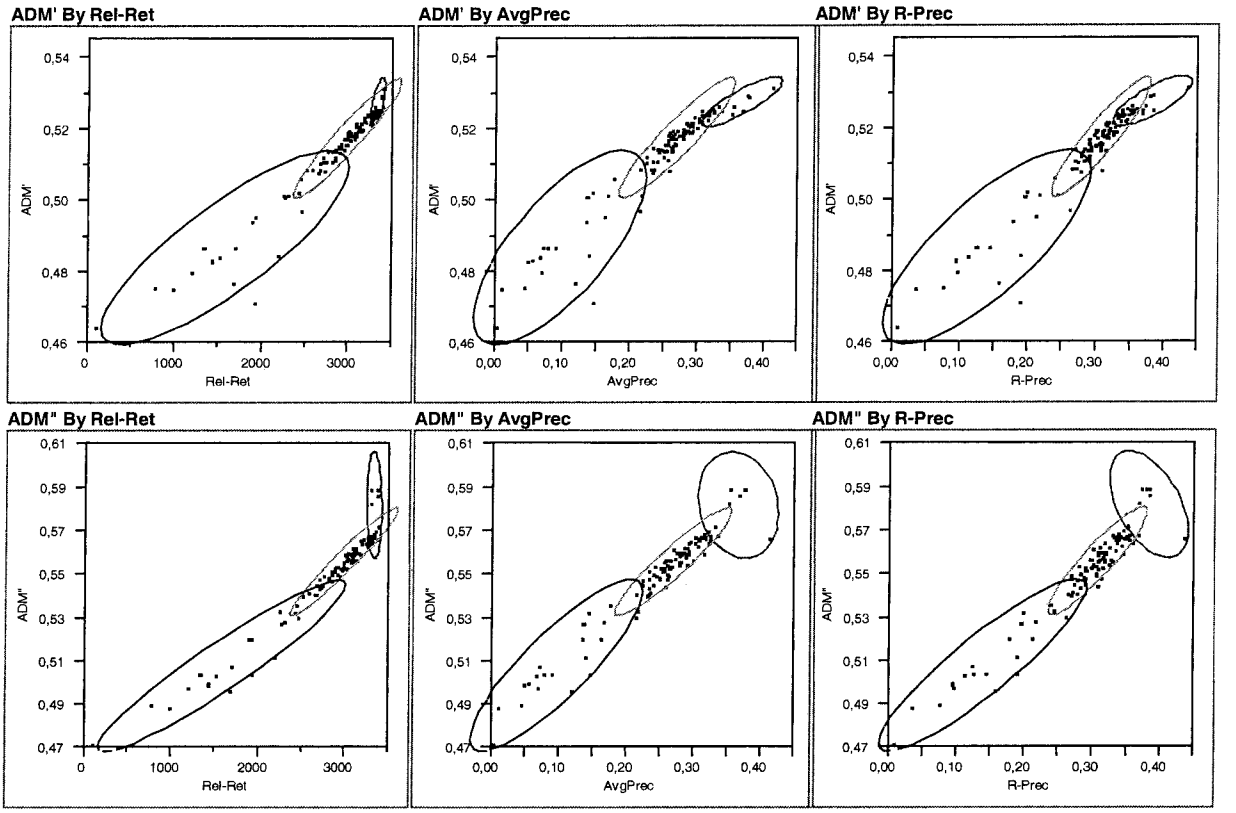

FIG. 7. Relationships among ADMs, reference measures, and effectiveness categories.

explained either by ADM not being able to capture all its effectiveness, or by ADM giving different effectiveness information if compared to that offered by the other measures.

Second Hypothesis

To test the second hypothesis, we evaluated ADM on seven subsets of the whole document set (i.e., relevant and

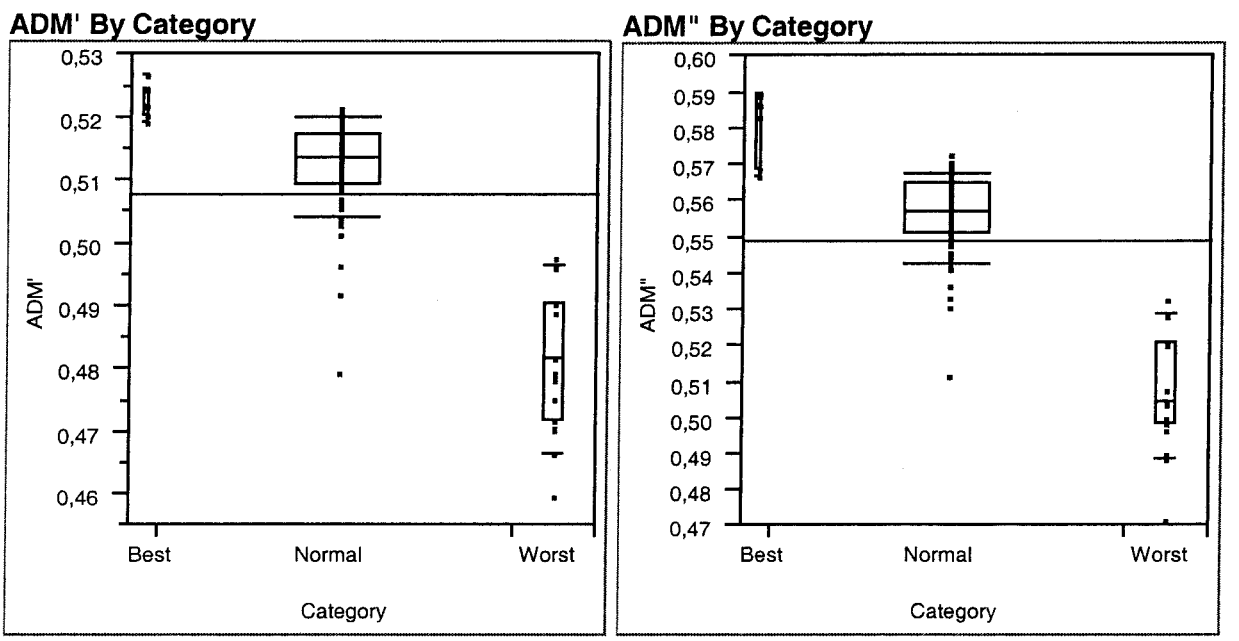

FIG. 8. Relationships between ADMs and effectiveness categories.

TABLE 4. ADMs Kendall correlations for subsets of data.

Set (Ret, Rel, topics)	No. of docs (approx.)	ADM′: Kendall's tau	ADM″: Kendall's tau
0 (100%, 100%, 100%)	53,000	1.000	1.000
1 (100%, 100%, 50%)	26,000	0.852	0.911
2 (100%, 100%, 20%)	10,000	0.674	0.765
3 (50%, 50%, 100%)	26,000	0.910	0.948
4 (10% 10% 100%)	5,000	0.802	0.870
5 (50% 50% 50%)	13,000	0.807	0.888
6 (100% 0% 100%)	50,000	0.935	0.976
7 (0% 100% 100%)	4728	0.846	0.886

retrieved documents, set 0 in the list below), obtained by choosing a random sample of relevant and retrieved documents, and/or selecting the first topics only:

0. 100% relevant, 100% retrieved documents, 100% of topics (whole document set);
1. 100% relevant, 100% retrieved, 50% of topics;
2. 100% relevant, 100% retrieved, 20% of topics;
3. 50% relevant, 50% retrieved, 100% topics;
4. 10% relevant, 10% retrieved, 100% topics;
5. 50% relevant, 50% retrieved, 50% topics;
6. Retrieved documents only, 100% topics;
7. Relevant documents only, 100% topics.

Average ADM′ and ADM″ obtained in this way have been then compared to the values obtained on the whole document set, by means of the Kendall correlation. Table 4 shows the results of the analysis; the number of documents in each set is approximated since each topic has a different number of relevant documents, each system retrieves a different subset of the relevant documents, and we made some random choices.

The data in Table 4 suggest that we might evaluate ADM on retrieved documents only (set 6), obtaining values that resemble very well those evaluated on the set given by relevant and retrieved documents, that is slightly larger and, more importantly, more difficult to be determined exactly, because of the well-known "dark matter problem in IR" (i.e., it is difficult to find all the relevant documents in the database) (Ingwersen, 1992; Zobel, 1998). Moreover, when the proportion of relevant documents is known to be high enough, as it is, for example, in TREC (Voorhees & Harman, 2000), set 3, i.e., 50% of the retrieved and (about) 50% of the relevant documents can also be used to compute ADM in a reliable way. Finally, set 1 (50% of the topics) is also interesting, since the correlation is high enough to reflect on diminishing the number of topics. These data are even more interesting since URSs are obtained on the basis of binary relevance judgments and rather approximate SRSs. With continuous values, ADM potential can be fully exploited and the results might be even more positive.

Fig. 9 shows a sample relationship between the values of ADM′ and ADM″ calculated on the whole set versus the values obtained on sets 1 and 3.

Limits of This Study

Since ADM full potential is exploited when SRS and URS are expressed on a continuous range, and since the data set we used for the experimentations did not have such values, we had to simulate them starting from the binary relevance assessments and from the system rankings. The study is thus constrained by available data.

URS″ was defined to have a continuous URS, but its distribution is particular, with two peaks close to 0 and 1 due to the strong relationship with the document's qrels; furthermore, it weakly depends on the SRSs of the best systems.

In addition to this, SRS values computed as we did (i.e., from document positions) are rather unfair to the IRSs, since the relevant documents are, on average, about 100 per query, whereas we forced the SRS value of the 100th document at 0.9, therefore higher than the value that the ideal IRS would have chosen (i.e., 0).

We believe that the way we simulated "real" URS and SRS values has hindered ADM's full potential. This can also explain, in part, why ADM values are in the range between 0.45 and 0.6, although another reason is the way ADM is calculated as a linear measure of differences (other distance models might give wider ranges).

Finally, at this stage we did not consider ADP and ADR.

Conclusions and Future Work

We have proposed ADM (average distance measure), a new measure of retrieval effectiveness based on continuous views of relevance and retrieval. A conceptual analysis shows that ADM could be an adequate replacement and improvement of standard effectiveness measures. Some experimental data support the conceptual analysis and, moreover, demonstrate that ADM can be a solution to the so-called "dark matter problem in IR" (Ingwersen, 1992), since the calculation of ADM on a subset of the relevant documents gives very similar results to those obtained by using all the relevant documents.

We plan to continue this research in several ways, with respect to both specific points and general issues. Let us see the specific points first. The conceptual comparison in the "Adequacy of ADM" section concerns mainly precision and recall, but it is easy to see that other measures that evaluate the rank provided by the IRS can also be criticized in similar ways (see Footnote 3). We also need to study the "linear scale assumption" issue, mentioned in Footnote 4. Experimental data are needed to understand if this assumption is a real problem or if its effects are negligible (it is important to understand, for example, if and how IRS comparisons are affected by the values assigned to the categories). Note, however, that the higher the number of categories, the less problematic the linear scale assumption is. A related issue is which methods to use for gathering continuous URSs (i.e., continuous relevance judgments). Although several studies seem to support the view that continuous relevance assess-

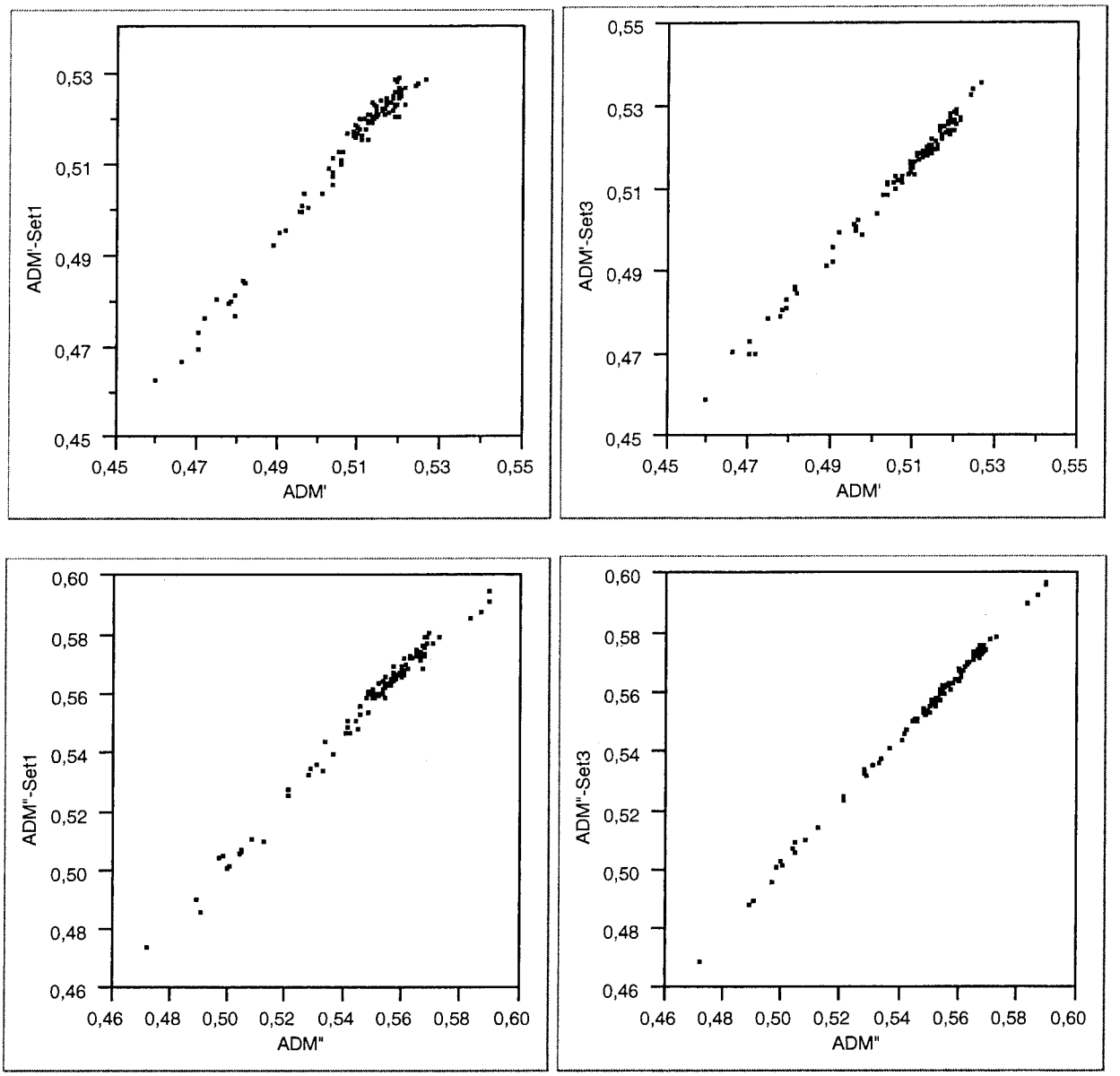

FIG. 9. Sample relationships between ADM values on the whole set and on sets 1 and 3.

ments are possible and reliable (Bruce, 1994; Janes, 1991b; Janes, 1994; Rorvig, 1988), more experimental evidence is needed.

We will also exploit other test-beds presenting both nonbinary URS values, as TREC-9 Web track (Voorhees, 2001) or other available databases, and reliable SRS values, like a subset of the systems participating in TREC. We will study the properties of ADP and ADR measures. We will use the systems scores in place of the rank (as we did in this work) as well. The easiest way to do that is to select those IRSs giving SRS values in the [0,1] range; we will also use the IRSs whose scores, though not in the [0,1] range, are consistent with the ranking, and normalize them. The normalization function is another interesting issue to investigate: a linear normalization (the simplest solution) from $[0,+\infty[$ to [0,1] seems not adequate, and other alternatives should be investigated. Moreover, starting from the distance measure concept discussed in the present work, further studies might involve different distance models (i.e., on the basis of quadratic measures of differences).

Turning to the general issues, we can study more foundational research problems. To discuss them, let us start by singling out some limitations that ADM, like the other effectiveness measures, exhibits. A first limitation is that ADM does not take into account the uncertainty in SRSs. To understand this point, let us consider two IR systems, IRS1 and IRS2, and let us assume that both IRS1 and IRS2 give the same 0.5 SRSs to a document *d*. However, IRS1 choice of this score is based on full knowledge that 0.5 is the most likely estimate of the URS of the document, since *d* contains evidence for both

relevance and nonrelevance, whereas IRS2 choice is based on a complete lack of knowledge, and IRS2 score is just the choice that maximizes ADM when no knowledge is available.[5] Now, IRS1 and IRS2 have the same ADM value, whereas the higher certainty of IRS1 should be rewarded by a higher effectiveness. Of course, there is the interesting, and open, issue of which IRS is more effective: one with a very certain, but slightly wrong, SRS, or one with a more uncertain, but more correct, SRS?

A second limitation is shown by another example.[6] Let us consider a situation in which URSs are either 0 or 1 (a truly binary relevance) and two IRSs, IRS1 and IRS2. IRS1 just sticks with the computed SRSs, whereas IRS2, derived from IRS1, exploits the binary relevance information, and "collapses" all its SRSs to either 0 or 1, depending on the 0.5 threshold. Now, if the two IRSs are "good" ones, IRS1 will have a higher ADM than IRS2, but IRS2 will be more convenient for the task of a user looking at the results of the two systems, since the SRSs can be used to rank the documents and read them in a more effective order. However, one might add that IRS2 is mixing, in some way, two different kinds of information: its estimation of URS (which is either 0 or 1) and its uncertainty on it. Moreover, the "ideal" IRS would give either 0 or 1 SRSs, with full certainty.

Of course, these limitations need a better understanding and further analysis. However, in our opinion, these open problems also show how the issue of IR effectiveness measures based on nonbinary relevance deserves much further analysis and how this article is just a preliminary work on a much longer research avenue. Actually, from a general point of view, the aim of this research should be to replace an estimate of the probability of (binary) relevance with an estimate of the amount of (in general, continuous) relevance. Let us state again that the probability of relevance provided by current probabilistic IRSs and the SRS (i.e., an estimate of the URS) that we would like to have are different in nature, even if both of them are a real number in [0,1]. Moreover, in an even more general way, URS and SRS values could be replaced by probability distributions, thus replacing the problem of estimating the amount of relevance with the (more difficult!) problem of fitting a probability distribution. If one agrees with the view proposed in this article, the Probability Ranking Principle (Robertson, 1977) is also questioned on its very foundations, since it assumes in an explicit way that the IRSs "response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness," and in an implicit but obvious way that one can speak of the "probability of relevance."

In other words, this research line should have the very ambitious indeed, threefold aim of: (i) replacing binary relevance with continuous relevance; (ii) replacing the estimate of the probability of relevance with the estimate of the amount of relevance; and (iii) replacing the amount of relevance with a probability distribution of relevance.

This is not an easy task at all, and it questions some of the fundamental assumptions on which IR researchers have based their work for decades: the assumption of binary relevance, the probabilistic model, the Probability Ranking Principle, etc. We are not alone in this effort, though. Manmatha, Rath, & Feng (2001) work on combining the output of different search engines by modelling the score distributions of IRSs, and therefore agree with our view that the score is important, not only the ranking. The same view will probably be endorsed by several researchers working on meta-search engines, an issue that is becoming of greater importance given the increasing circulation of mobile devices and of peer-to-peer computing. The combination of the results of different IRSs should be more effective when using the SRSs, rather than using the ranking alone, since more information is available. The Probability Ranking Principle, or rather the assumptions on which it is based, have been questioned for years (see Cooper, 1994). The ranked list of retrieved documents is sometimes seen as "the QWERTY of information retrieval"[7] and, nowadays, several IRSs show the search results in a different way (e.g., http://www.kartoo.com/). Finally, IR models based on continuous relevance are not available at the moment, whereas they should be developed, tested, and compared with the current IR models based on binary relevance. Perhaps these new "continuous" models can be obtained simply by extending the "binary" ones (e.g., by using the probabilistic IR model to obtain a relevance score starting from the probability of binary relevance), or perhaps completely new theoretical tools need to be exploited. This leaves plenty of space for future work.

Acknowledgments

We would like to thank the TREC organizers who kindly provided us with the TREC-8 data. We are particularly grateful to Stephen Robertson for long, detailed, and stimulating criticisms on a previous version of this article; we regret that we have not been able to take all of them into account. We also thank the referees who provided useful suggestions.

References

Bookstein, A. (1979). Relevance. Journal of the American Society for Information Science, 30, 269–273.

[5]The reader will note the similarity between this example and the well-known example of a Ming vase that contains evidence both for and against its authenticity, and whose authenticity is judged by an expert and by an ignorant.

[6]This example has been suggested by Steve Robertson.

[7]We have heard this expression from Matt Jones. It is well known by ergonomics studies that "QWERTY" keyboards are not the most effective solutions to text typing, since keyboards with different key distributions allow higher typing speed. However, their usage is so widespread that it turns out to be impossible to change this de facto standard.

Borlund, P., & Ingwersen, P. (1998). Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. Proceedings of the 21st Annual International ACM SIGIR Conference (pp. 324–331). Melbourne, Australia.

Bruce, H.W. (1994). A cognitive view of the situational dynamism of user-centered relevance estimation. Journal of the American Society for Information Science, 45, 142–148.

Buckley, C., & Voorhees, E.M. (2000). Evaluating evaluation measure stability. Proceedings of the 23rd Annual International ACM SIGIR Conference (pp. 33–40). Athens, Greece.

Buckley, C., & Voorhees, E.M. (2002). The effect of topic set size on retrieval experiment error. Proceedings of the 25th Annual International ACM SIGIR Conference (pp. 316–323). Tampere, Finland.

Burgin, R. (1999). The Monte Carlo method and the evaluation of retrieval system performance. Journal of the American Society for Information Science, 50, 181–191.

Cooper, W.S. (1994). The formalism of probability theory in IR: A foundation or an encumbrance? Proceedings of the 17th Annual International ACM SIGIR Conference (pp. 243–247). Dublin, Ireland.

Eisenberg, M. (1986). Magnitude estimation and the measurement of relevance. Unpublished doctoral thesis, Syracuse University, Syracuse, NY.

Eisenberg, M., & Hu, X. (1987). Dichotomous relevance judgments and the evaluation of information systems. Proceedings of the American Society for Information Science (pp. 66–69). Medford, NJ: Learned Information.

Eisenberg, M.B. (1988). Measuring relevance judgments. Information Processing & Management, 24, 373–389.

Frei, H., & Schauble, P. (1991). Determining the effectiveness of retrieval algorithms. Information Processing & Management, 27, 153–164.

Harter, S.P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47, 37–49.

Ingwersen, P. (1992). Information retrieval interaction. London: Taylor Graham.

Janes, J.W. (1991a). The binary nature of continuous relevance judgments: A case study of users' perceptions. Journal of the American Society for Information Science, 42, 754–756.

Janes, J.W. (1991b). Relevance judgments and the incremental presentation of document representations. Information Processing & Management, 27, 629–646.

Janes, J.W. (1994). Other people's judgments: A comparison of user's and other's judgments of document relevance, topicality, and utility. Journal of the American Society for Information Science, 45, 160–171.

Janes, J.W., & McKinney, R. (1992). Relevance judgments of actual users and secondary judges. Library Quarterly, 62, 150–168.

Järvelin, K., & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. Proceedings of the 23rd Annual International ACM SIGIR Conference (pp. 41–48). Athens, Greece.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems (TOIS), 20, 422–446.

Korfhage, R.R. (1997). Information storage and retrieval. New York: John Wiley & Sons.

Losee, R.M. (2000). When information retrieval measures agree about the relative quality of document rankings. Journal of the American Society for Information Science, 51, 834–840.

Manmatha, R., Rath, T., & Feng, F. (2001). Modeling score distribution for combining the outputs of search engines. Proceedings of the 24th Annual International ACM SIGIR Conference, 267–275. New Orleans, LA.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48, 810–832.

Mizzaro, S. (1998). How many relevances in information retrieval? Interacting with Computers, 10, 305–322.

Mizzaro, S. (2001). A new measure of retrieval effectiveness (Or: What's wrong with precision and recall). In T. Ojala (Ed.), International Workshop on Information Retrieval (IR'2001) (pp. 43–52). Infotech Oulu, Oulu, Finland.

Robertson, S.E. (1977). The probability ranking principle in IR. Journal of Documentation, 33, 294–304. Republished in K. Sparck Jones, & P. Willett. (1997), Readings in information retrieval (pp. 281–286). San Francisco, CA: Morgan Kaufmann.

Rorvig, M.E. (1988). Psychometric measurement and information retrieval. Annual Review of Information Science and Technology, 23, 157–189.

Salton, G., & McGill, M.J. (1984). Introduction to modern information retrieval. New York: McGraw-Hill.

Schamber, L. (1994). Relevance and information behavior. Annual Review of Information Science and Technology, 29, 3–48.

Schamber, L., Eisenberg, M.B., & Nilan, M.S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26, 755–776.

Sormunen, E. (2002). Liberal relevance criteria of TREC—Counting on negligible documents? In M. Beaulieu et al. (Eds.), Proceedings of the 25th Annual ACM SIGIR Conference (pp. 324–330). Tampere, Finland.

Sparck Jones, K. (1974). Automatic indexing. Journal of Documentation, 30, 393–432.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. Information Processing & Management, 34, 599–621.

Swets, J.A. (1967). Effectiveness of information retrieval methods. Cambridge, MA: Bolt, Beranek and Newman.

van Rijsbergen, C.J. (1979). Information retrieval (2nd ed.). London: Butterworths.

Voorhees, E.M. (2001). Evaluation by highly relevant documents. Proceedings of the 24th Annual International ACM SIGIR Conference (pp. 74–82). New Orleans, LA.

Voorhees, E.M., & Harman, D. (2000). Overview of the Eighth Text Retrieval Conference (TREC-8), The 8th Text Retrieval Conference (TREC-8) (pp. 1–24) (NIST SP-500-246). Available: http://trec.nist.gov/

Yao, Y.Y. (1995). Measuring retrieval effectiveness based on user preference of documents. Journal of the American Society for Information Science, 46, 133–145.

Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? Proceedings of the 21st Annual International ACM SIGIR Conference (pp. 307–314). Melbourne, Australia.